INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Practical Construction Against Theoretical Approach in Fingerprinting

Fabien Galand

N° 6080

décembre 2006

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--1999--FR+ENG

INRIA
RENNES

Practical Construction Against Theoretical Approach in Fingerprinting

Fabien Galand *

Thème COG — Systèmes cognitifs
Projet TEMICS

Rapport de recherche n° 6080 — décembre 2006 — 13 pages

Abstract: We consider fingerprinting under collusion attacks in the Hamming space, using the framework of [SBM05]. We construct a family of fingerprinting codes efficient against coalition of arbitrary size: Using this family, tracing dishonest users can be done without error and in polynomial time. The number of users is exponential in the length of the code. The proposed construction relies on centered error correcting codes [BP99] for which we discuss two constructions. Our results have an amazing relation with an upper bound on the number of users derived in [SBM05]: dropping two assumptions we construct codes beating their bound, still keeping practical properties.

Key-words: bound, capacity, centered error correcting codes, construction, covering codes, fingerprinting, random codes, Reed-Solomon codes, marking assumption, distortion assumption.

* fabien.galand@irisa.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

Construction explicite et approche théorique en fingerprinting

Résumé : Nous étudions les attaques par collusion dans l'espace de Hamming en utilisant le cadre de [SBM05]. Nous construisons une famille de codes anti-collusions efficace contre des coalitions de taille arbitraire : cette famille permet un traçage sans erreur des utilisateurs malhonnêtes et cela en temps polynomial. Le nombre d'utilisateurs est exponentiel en la longueur. La construction proposée repose sur les codes correcteurs d'erreurs centrés [BP99] pour lesquels nous étudions deux constructions. Nos résultats entretiennent une relation étrange avec une borne supérieure sur le nombre d'utilisateurs obtenue dans [SBM05] pour les codes à composition constante : les codes que nous construisons, sans être à composition constante, dépassent cette borne.

Mots-clés : borne, capacité, codes correcteurs d'erreurs centrés, construction, codes couvrants, fingerprinting, codes aléatoires, codes de Reed-Solomon, marking assumption, distortion assumption.

1 Introduction

When distributing digital contents, one may want to prohibit redistribution of the contents, for example when one sells video. A reasonable requirement to enforce this prohibition is to be able to trace users have redistibuted their copies.

This can be achieved in a simple way if users do not collude, and just redistribute their own original copies : the dealer can embed, in a robust way, a kind of serial number in each copy. Watermarking techniques address this embedding issue. Fighting collusion of users is more involved.

Since the introduction of this topic by [BMP85], the most studied framework is the so-called *marking assumption* popularized by [BS98]. In this setting, the extra data is a collection of marks, that is a n-tuple over some alphabet Σ. The marking assumption states that, possibly after being proceeded by some decoder, a fake copy created by some colluders will contain a n-tuple $z = (z_0, ..., z_{n-1})$ such that in each position $i \in \{0, ..., n-1\}$, at least one of the colluder has symbol z_i in position i.

A more recently investigated model is the one used in [SBM05]. Here, the marking assumption is replaced by two constraints. The first is related to the creation of user copies : the copies must be close to the original content. The second concernes the possible fakes : a fake copy can be created by a coalition of dishonest users if and only if the fake copy is close enough to, at least, one of the members' copies. For these reasons, we will call this model the *distortion assumption*.

In those models, several problems arise. Among them, there are

1. the effectiveness of the solution proposed, that is the possibility to create the user copies and to trace dishonest users from fake copies using polynomial time algorithms in the content length;

2. the scalability, that is, essentially, the asymptotic behaviour of the user rate in function of the other parameters.

In this work, we will address these two problems in the settings of [SBM05], namely what we call the distortion assumption. We will see, that in a very surprising fashion, it is possible to tune the parameters of the fingerprinting systems in such a way that we avoid any benefit of collusion. This is very different from what can be done with the marking assumption (e.g. [BBK03]).

In [SBM05], the maximum number of users is investigated, under the distortion assumption and in a non constructive way, for a family of fingerprinting systems. It is proved that the logarithm of this maximum scales with $\mathcal{O}(1/L)$ and goes to zero if L grows linearly with n.

In this paper, we construct a family of binary fingerprinting codes with exponentially many users (in the content length). This family allows efficient user copy creation and dishonest user tracing : both of these issues can be solved with polynomial time algorithm. Moreover, our fingerprinting codes have a very interesting new property: they are effective against coalitions of arbitrary size. This, with the exponential number of users, implies that our codes beat the bound derived in [SBM05] and, so, that the family of fingerprinting systems considered in [SBM05] is inefficient.

To achieve efficent fingerprinting codes, we heavily rely on centered error correcting codes. In fact, we prove that fingerprinting for collusion attack under the distortion assumption is a generalisation of error correction for centered codes. Thus, we address fingerprinting via the particular problem of centered codes.

The paper is organised as follow. In Section 2 we present the framework for fingerprinting codes under the distortion assumption and our notation. Section 3 gives basic facts about the centered error correcting codes, and states a simple upper bound. We address the construction of centered codes in Section 4: First, in Section 4.1 we consider (unpractical) random construction to obtain lower bound and in Section 4.2 we give a construction which is fully practical. Section 5 explains how to use centered error correcting codes to construct fingerprinting codes. We also show how this construction allows to fight coalition of arbitrary size. We detail the parameters achieved when using the centered codes constructed in Section 4.2. Section 6 briefly compares

our results and others results on the capacity of some fingerprinting systems obtained in [SBM05]. Finally, we conclude pointing out an important difference between the model we use here and the marking assumption.

2 Fingerprinting Formal Problem

We have a set $\mathcal{V} \subset \mathbb{F}^n$ of original data, which can be seen as sequences of length n over some alphabet $\mathbb{F}$. In order to fingerprint some copyright protected data v, the dealer is allowed to change at most Δ_o coordinates, that is a fingerprinted copy of v is in $\mathrm{B}_{\Delta_o}(v) = \{y : \mathrm{d}(v, y) \leq \Delta_o\}$ — d is the Hamming distance over $\mathbb{F}^n$ —, the ball of center v and radius Δ_o.

Definition 2.1 *An (n, M, Δ_o) fingerprinting code is a mapping, $E : \mathcal{V} \times \mathcal{M} \to \mathbb{F}^n$, such that*

1. *$|\mathcal{M}| = M$;*
2. *$E_v : m \mapsto E(v, m)$ is injective for all $v \in \mathcal{V}$, i.e.*
$$\forall (m, m') \in \mathcal{M}^2 \qquad E_v(m) = E_v(m') \Longrightarrow m = m' \ ;$$
3. *$\forall v \in \mathbb{F}^n \quad E(v, \mathcal{M}) \subset \mathrm{B}_{\Delta_o}(v)$ ($E(v, \mathcal{M})$ being the set of all possible fingerprinted copies of v)*

A coalition $U \subset E(v, \mathcal{M})$ is a set of fingerprinted copies of some $v \in \mathcal{V}$, and a forgery f created by this coalition must satisfy :
$$\exists c \in U \quad \mathrm{d}(c, f) \leq \Delta_f \ ,$$
for some fixed value Δ_f. Thus, the set of forgeries computable from a set U is
$$\mathrm{forg}_{\Delta_f}(U) = \bigcup_{c \in U} \mathrm{B}_{\Delta_f}(c) \ .$$

Definition 2.2 *A fingerprinting code is said to be (L, Δ_f)-secure if there exists a mapping $D : \mathbb{F}^n \to \mathbb{F}^n$ (called the tracing or decoding mapping), such that for any $v \in \mathbb{F}^n$ and any coalition $U \subset E(v, \mathcal{M})$ of size at most L,*
$$D\left(\mathrm{forg}_{\Delta_f}(U)\right) \subset \widehat{U} \ ,$$
where $\widehat{U}$ is the set of messages corresponding to the copies in U, $\widehat{U} = \{m \mid E(v, m) \in U\}$.

Remark that this definition of secure fingerprinting codes do not allow any kind of probability of error, neither a failure in the decoding (no output), neither to frame an innocent user. Moreover, we do not require any knowledge of the original data v.

In the sequel, we will set $\mathcal{V} = \mathbb{F}^n$, with $\mathbb{F}$ the finite field with 2 elements. The number of users M will be a power of 2, $M = 2^\ell$. The set $\mathcal{M}$ will be $[1, M]$, and when it will be convenient, we will identify an integer with its binary expansion, and $\mathcal{M}$ with $\mathbb{F}^\ell$.

3 Centered Error Correcting Codes

Centered error correcting (CEC) codes, which were introduced in [BP99], are a generalization of the classical notion of error correcting codes.

Roughly speaking, we add a condition on the localization of the codeword c encoding a message : we have a pair (v, m) and c must be within the ball $\mathrm{B}_T(v)$ of center v and radius T, where T is a new parameter of the code.

Definition 3.1 *A centered error correcting code of parameters $(n, M, T, 2e + 1)$ is defined by an encoding mapping $E : \mathbb{F}^n \times \mathcal{M} \to \mathbb{F}^n$ such that*

1. $\forall (v,m)$ *and* $(v',m') \neq (v,m)$,

$$\mathrm{B}_e(E(v,m)) \bigcap \mathrm{B}_e(E(v',m')) = \emptyset$$

2. $\forall (v,m), \quad E(v,m) \in \mathrm{B}_T(v)$

The first condition is a classical one to allow correction of e Hamming errors. The second condition is the localization constraint.

As easily seen, setting $T = n$ leads to the usual definition of error correcting codes since the condition 2) of the above definition is then always satisfied.

Another way, equivalent to Definition 3.1, to define CEC codes is by mean of coverings.

Proposition 3.2 *A centered error correcting code of parameters* $(n, M, T, 2e+1)$ *is a set* $\{C_i\}$ *of* M *disjoint coverings of radius* T *of the space* $\mathbb{F}^n$*, distant from each other of at least* $2e+1$ *:*

1. $\forall i, \quad \bigcup_{c \in C_i} B_T(c) = \mathbb{F}^n$;

2. $\forall i, j \neq i\,, \forall (c,c') \in C_i \times C_j, \quad \mathrm{d}(c,c') \geq 2e+1.$

Proof. The sets $E(\mathbb{F}^n, m)$, where $m \in \mathcal{M}$, are coverings of radius T : for all v, $E(v,m)$ is at distance at most T from v by 2) of Definition 3.1. Moreover, those sets are $2e+1$ apart from each other: otherwise, there exist couples (v,m) and $(v',m') \neq (v,m)$ with $\mathrm{B}_e(E(v,m)) \bigcap \mathrm{B}_e(E(v',m')) \neq \emptyset$, which would contradict 1) of Definition 3.1. The reverse way can be proved as easily as this one. □

3.1 On Non-Existence Results

The following lemma, which can by found in [BP99] allows to derived an upper bound on the cardinality M of a CEC code.

Lemma 3.3 *Let* T *and* e *be non negative integers, with* $T \geq e$*. Denote by* $\mathcal{N}_a(Y)$ *the set of binary words at distance at most* a *from the set* Y

$$\mathcal{N}_a(Y) = \bigcup_{y \in Y} B_a(y) \ .$$

Then, for all set Y*, the following inequality holds*

$$\left| \frac{\mathcal{N}_T(Y)}{\mathcal{N}_e(Y)} \right| \leq \left| \frac{V_T}{V_e} \right| \ ,$$

where V_a *is the cardinality of a ball of radius* a*.*

Consider a CEC code of parameters $(n, M, T, 2e+1)$ with encoding mapping E. We will apply this lemma to the set

$$Y(m) = \{c \mid \exists v \quad E(v,m) = c\} = E(\mathbb{F}^n, m)$$

of codewords that encode some message m. The set $Y(m)$ is a covering of radius T, thus

$$|\mathcal{N}_T(Y(m)) = 2^n| \ .$$

On the other hand, since the balls of radius e centered on the elements of $Y(m)$ are disjoint, we have

$$\begin{aligned} M &\leq \frac{2^n}{\mathcal{N}_e(Y(m))} \\ &\leq \frac{\mathcal{N}_T(Y(m))}{\mathcal{N}_e(Y(m))} \ . \end{aligned}$$

Lemma 3.3 yields

Proposition 3.4 *For all CEC codes of parameters $(n, M, T, 2e+1)$, we have the upper bound*

$$\log(M) \leq \log(V_T) - \log(V_e) \ .$$

The asymptotic behavior of the volume of the ball is well-known:

1. when the radius a grows linearly with the length n we have

$$\log(V_a) = n \cdot h\left(\frac{a}{n}\right) + o(n) \ ,$$

where $h(x) = -x\log(x) - (1-x)\log(1-x)$ is the usual binary entropy (e.g. see [MS96, chap. 10, §11, Cor. 9])

2. when the radius a is fixed and n grows to infinity,

$$\log(V_a) = a\log n - \log(a!) + o(1) \ .$$

We have therefore the following two asymptotic upper bounds

Corollary 3.5 *For all $(n, M, T, 2e+1)$ CEC codes of length n large enough, we have*

1. *For $\tau = T/n$ and $\varepsilon = e/n$ fixed,*

$$\log(M) \lesssim n \cdot (h(\tau) - h(\varepsilon)) \ .$$

2. *For T and e fixed,*

$$\log(M) \lesssim (T-e) \cdot \log(n) - \log\left(\frac{T!}{e!}\right) \ .$$

where we write $f \lesssim g$ for $f \leq g(1+o(1))$ when n tends to infinity.

4 Constructing CEC Codes

4.1 On Existence Results

A simple way to construct CEC codes is to use linear codes.

Proposition 4.1 *Let C be a $[n, k, 2e+1]$ binary linear code. Let $C' \subset C$ be a subcode of dimension k' and covering radius T. The cosets of C' in C,* i.e. *the sets $x + C'$, $x \in C$, form a CEC code of parameters $(n, 2^{k-k'}, T, 2e+1)$.*

Proof. The code C can be partitioned in $2^{k-k'}$ cosets of C', the cosets are at distance at least $2e+1$ from each other since they are composed of codewords of C, and these cosets have the same covering radius as C', namely T. This is exactly the definition (in fact the equivalent one given in Proposition 3.2) of a $(n, 2^{k-k'}, T, 2e+1)$ CEC code. □

4.1.1 Encoding

Recall that, for encoding with a CEC code, each of the $M = 2^{k-k'}$ sets C_i is used to encode a particular message $m \in \{1, ..., M\}$, *i.e.* a word c encodes the message m if and only if $c \in C_m$. Moreover, a codeword encoding the message m for a center v can be found at distance at most T, since T is the covering radius of each set C_i. Now, consider a parity-check matrix H of C. A parity-check matrix H' of C' can be obtained by adding rows to H,

$$H' = \begin{pmatrix} H \\ Q \end{pmatrix} \tag{1}$$

for some $(k-k') \times n$ matrix Q. The cosets C_i of C' in C can be indexed in such a way that

$$C_i = \{c \in \mathbb{F}^n \mid c \cdot H = 0 \text{ and } c \cdot Q = i\}$$

(where needed, the integer i is identified with its binary expansion). So, we are looking for a word $E(v,m)$ such that

$$\begin{aligned} E(v,m) \cdot \begin{pmatrix} H \\ Q \end{pmatrix}^t &= \begin{pmatrix} E(v,m) \cdot H^t & E(v,m) \cdot Q^t \end{pmatrix} \\ &= (0 \ m) \ . \end{aligned}$$

We can choose $E(v,m) = v + u$ where u is a solution of $u \cdot (H^t \ Q^t) = (0 \ m)$ with $w(u) \leq T$.

4.1.2 Decoding

If a binary word z of weight at most e is added to a codeword $E(v,m)$ to produce y, since $E(v,m)$ is a codeword of the e-error correcting code C, we can recover $E(v,m)$ by adding to y the (unique) word of weight at most e and syndrome $y \cdot H^t$, namely z. Then, we retrieve the message by computing $(y+z) \cdot Q^t$.

4.1.3 An Asymptotic Lower Bound

When $\tau = T/n$ and $\varepsilon = e/n$ are fixed and n grows to infinity, two classical results on random binary linear codes allow to derive the asymptotic behavior of CEC codes constructed in Proposition 4.1. We sum up these classical results in the following theorem and refer to [Bar98, Lem. 1.2 and Th. 3.4] for detailed proofs.

Theorem 4.2 *Let α be a real number less than $1/2$ and let n grow to infinity. A $[n, n \cdot (1-h(\alpha))]$ linear code has minimum distance at least*

$$n \cdot \alpha$$

and covering radius

$$n \cdot \alpha \cdot (1 + o(1))$$

with probability tending to 1.

Let C be a $[n, n \cdot (1-h(2\varepsilon))]$ code. It has minimum distance at least $2 \cdot n \cdot \varepsilon$ with hight probability and thus corrects errors of weight up to $n \cdot \varepsilon$. Let $C' \subset C$ be a subcode of dimension $n \cdot (1-h(\tau))$ (this implies $2\varepsilon < \tau$): it has a covering radius $n \cdot \tau$ with hight probability. Applying the construction detailed at the beginning of this section leads to a CEC code with parameters

$$(n, 2^{n(h(\tau)-h(2\varepsilon))}, n \cdot \tau, 2 \cdot n \cdot \varepsilon) \ .$$

Proposition 4.3 *Providing that $\tau > 2\varepsilon$, the construction of Proposition 4.1 allows to achieve centered error correcting codes with parameters*

$$(n, M, n \cdot \tau, 2 \cdot n \cdot \varepsilon) \ ,$$

where

$$\log M \gtrsim n \cdot (h(\tau) - h(2\varepsilon)) \ .$$

The last proposition is to be compared to the corresponding upper bound stated in 1) of Corollary 3.5.

4.1.4 Effectiveness

Centered error-correcting codes constructed with random codes have two main flaws that prevent them from being usable in practice. First, we cannot verify the "assumed" parameters. In fact, it is known by [Var97] and [McL84] that, for linear codes, the decision problems associated with minimum distance and covering radius computations are respectively NP-complete and Π_2-complete (the Π_2 complexity class includes the NP class). Moreover, even efficient probabilistic algorithms are not known. Second and even more painful, both encoding and decoding need to find words z of weight at most respectively T and e with a particular syndrome with respect to a linear code (in fact, decoding up to e for C and complete decoding for C'). Unfortunately, this is an NP-hard problem and, once more, we do not even know efficient probabilistic algorithm (e.g. see [Bar98]).

Nevertheless, using specific linear codes, this problem may be overcome. For instance, writing BCH(e) the (narrow sense and primitive) binary e-error-correcting BCH code (see [MS96, Chap. 7, §6]), we can choose $C = \text{BCH}(e)$ and $C' = \text{BCH}(e')$ with $e' > e$, since those codes are nested, *i.e.* C' is a subcode of C. For the choice (e, e') equal to $(1, 2)$, $(1, 3)$ and $(2, 3)$, the construction is effective since complete decoding are known (see [Ber68] and [vHB76]). The corresponding CEC codes have respectively the parameters $(2^r - 1, 2^r, 3, 3)$, $(2^r - 1, 2^{2r}, 5, 3)$ and $(2^r - 1, 2^r, 5, 5)$ (see [MS96, Chap. 9] and [CHLL97, Chap. 10] for the parameters of BCH codes).

A nice example of encoding and decoding, with $C = \text{BCH}(1)$, $C' = BCH(2)$ and $r = 4$, can be found in [ZC91, §IV.C], where a construction similar to the one we presented in this section (with the restriction $e = 1$) is described and applied to some BCH codes.

4.2 An Explicit Construction

The construction we propose relies on two codes: a direct sum of T binary Hamming codes of length $2^r - 1$ and an e-error-correcting code over a larger alphabet, the finite field with 2^r elements, denoted $\mathbb{F}_{2^r}$. For the purpose of this construction, we will need to identify elements of vector space $\mathbb{F}^r$ and elements of the field $\mathbb{F}_{2^r}$. To do so, we choose a fixed basis of $\mathbb{F}_{2^r}$ over $\mathbb{F}$, and write $\overline{x}$ the element in $\mathbb{F}_{2^r}$ corresponding to $x \in \mathbb{F}^r$. With a slight abuse in notation, we also write $\overline{c}$ the element of $\mathbb{F}_{2^r}^T$ corresponding to $c \in \mathbb{F}^{rT} \simeq \mathbb{F}^r \times ... \times \mathbb{F}^r$, where the identification is done componentwise.

We will use the definition of CEC codes given by the Proposition 3.2 in the proof of the following result

Theorem 4.4 *Let* $\mathbf{C}$ *be a* $[T, k, 2e+1]$ *code over* $\mathbb{F}_{2^r}$*. There exists a binary* $((2^r-1)\cdot T, 2^{k\cdot r}, T, 2e+1)$ *centered code* C*. Moreover,*

1. *if* $\mathbf{C}$ *has a polynomial time encoding algorithm, so has* C*;*
2. *if* $\mathbf{C}$ *has a polynomial time decoding algorithm, so has* C*.*

In the above theorem, the algorithms are polynomial with respect to the corresponding code length, that is T for $\mathbf{C}$ and $n = (2^r - 1) \cdot T$ for C.

Proof. Let $\mathcal{C}$ be the direct sum of T Hamming codes of length $2^r - 1$, thus the length n of $\mathcal{C}$ is $n = (2^r - 1) \cdot T$, and let H be a parity check matrix of a Hamming code of length $2^r - 1$. A parity check matrix of $\mathcal{C}$ can be obtained from H as a block matrix by

$$P = \begin{pmatrix} H & & 0 \\ & \ddots & \\ 0 & & H \end{pmatrix} . \tag{2}$$

We can now define our CEC code C, as the following set of cosets of $\mathcal{C}$

$$\left\{ y + \mathcal{C} : \overline{y \cdot P^t} \in \mathbf{C} \right\}$$

where P^t is the transpose of P. Since cosets have the same covering radius as the code $\mathcal{C}$, and $\mathcal{C}$ has clearly a radius equal to T, we just have to prove that $y+\mathcal{C}$ and $z+\mathcal{C}$ are $2e+1$ apart when $y \cdot P^t \neq z \cdot P^t$. If we have $y' = y + c_1$ and $z' = z + c_2$, where $c_1, c_2 \in \mathcal{C}$, then

$$\overline{(y'+z') \cdot P^t} = \overline{(y+z) \cdot P^t} \in \mathbf{C} \setminus \{0\} \ .$$

But, since P is a block matrix, grouping the coordinates of y and z by $2^r - 1$, that is, writing $y = y_1 \dots y_T$ and $z = z_1 \dots z_T$, for some y_i,$z_i \in \mathbb{F}^{2^r-1}$, we have

$$\overline{(y+z) \cdot P^t} = \left(\overline{(y_1+z_1) \cdot H^t}, \dots, \overline{(y_T+z_T) \cdot H^t}\right) \ .$$

On the other hand, $\mathbf{C}$ has minimal distance $2e+1$. Thus, there exist at least $2e+1$ coordinates of $\overline{(y+z) \cdot P^t}$ that are different from 0. Equivalently, there exist $2e+1$ values of i such that $(y_i+z_i) \cdot H^t = (y'_i + z'_i) \cdot H^t \neq 0$, *i.e.* such that $\mathrm{d}(y'_i, z'_i) \geq 1$. Finally, we get

$$\mathrm{d}(y', z') = \sum_{0 \leq j < T} \mathrm{d}(y'_j, z'_j) \geq 2e+1 \ ,$$

completing the proof of the first part of the theorem.

To prove 1), we consider a word v and a message m. We use $\mathbf{C}$ to encode m in a codeword $\overline{c_m} \in \mathbf{C}$. Then, we look for a word $E(v,m)$ at distance at most T from v in the coset of $\mathcal{C}$ of syndrome c_m (with respect to P). This can be done by computing a word u of weight at most T and syndrome $c_m + v \cdot P^t$ and taking $E(v,m) = v + u$, since

$$\begin{aligned} E(v,m) \cdot P^t = (v+u) \cdot P^t &= v \cdot P^t + c_m + v \cdot P^t \\ &= c_m \ . \end{aligned}$$

The computational cost is an encoding in $\mathbf{C}$ and the search of u. For a direct sum of T Hamming codes of length $2^r - 1$, the computation of u consists in T decoding of a Hamming code of length $2^r - 1$, so it is $\mathcal{O}(rT)$ operations. Thus, if the encoding of m in a codeword of $\mathbf{C}$ can be done in polynomial time, the overall computational cost of $E(v,m)$ is polynomial.

To prove 2), we consider a word y at distance at most e from $E(v,m)$. We can recover m in the following way. First, compute the syndrome $s = y \cdot P^t$. Since $\mathrm{d}(y, E(v,m)) \leq e$, $\overline{s}$ is at distance at most e from $\overline{E(v,m) \cdot P^t}$. But $\overline{E(v,m) \cdot P^t}$ is a codeword of $\mathbf{C}$, and $\mathbf{C}$ can correct up to e errors. So, performing a decoding on $\overline{s}$ allows to recover m. The total cost is T syndrome computations with respect to H, and a decoding in $\mathbf{C}$. So, it is polynomial if $\mathbf{C}$ can be decoded in polynomial time. □

Remark that, in contrast to the construction of Proposition 4.1, the decoding used for the construction do not recover the codeword $E(v,m)$, but directly outputs the message m. Nevertheless, it is possible to compute $E(v,m)$: the codeword $E(v,m)$ is the unique codeword in $\mathrm{B}_e(y)$, since, by hypothesis, $y \in \mathrm{B}_e(E(v,m))$ and, by 1) of Definition 3.1, $\mathrm{B}_e(E(v,m)) \cup \mathrm{B}_e(E(v',m')) = \emptyset$ for $(v',m') \neq (v,m)$. Thus, there exists a unique word u, of weight at most e, such that $y + u = E(v,m)$. But $u \cdot P^t = c_m + y \cdot P^t$. Since we know m and y, we can compute right hand side of the last equation. Finding the solution u of weight at most e is then possible, thanks to the structure of P. Moreover, we can see that this is equivalent to the computation required by an encoding in the centered code. So, if $\mathbf{C}$ has a polynomial time encoding, recovering $E(v,m)$ can be done in polynomial time.

Corollary 4.5 *For $2^r > T > 2e$, there exist binary $((2^r-1)T, 2^{r(T-2e)}, T, 2e+1)$ CEC codes with polynomial time encoding and decoding algorithms.*

Proof. Since $q = 2^r$ is greater than T, we can use for $\mathbf{C}$ a $[T, T-2e, 2e+1]$ Reed-Solomon code over $\mathbb{F}_{2^r}$. □

The previous result gives us fully practical CEC codes. On the other hand, in view of asymptotic analysis, it has a drawback: T has to be less than 2^r. But, if we let the length $n = T(2^r - 1)$ go to infinity with $T \propto n$, this implies a fixed r and, so, a bounded T, and we have a contradiction.

Using algebraic geometric codes [TV91] for $\mathbf{C}$ allows to drop the upper bound condition on T. So, we can fix $q = 2^r$ (r must be even), and still have the length n of the CEC code going to infinity. From [TVZ82], we can have, asymptotically, for $\mathbf{C}$

$$\frac{k}{T} \leq 1 - \frac{2e+1}{T} - \frac{1}{\sqrt{q}-1}$$

for e/T fixed. Thus we can state another corollary to Theorem 4.4.

Corollary 4.6 *Let r be an even integer. Let $\tau = (2^r-1)^{-1}$ and ε such that $\tau > 2\varepsilon$. For length n large enough, there exist $(n, M, n\cdot\tau, 2\cdot n\cdot\varepsilon)$ CEC codes, with*

$$\log(M) \gtrsim n \cdot \left(\tau - 2\varepsilon - \frac{\tau}{2^{\frac{r}{2}}-1}\right) \cdot \log\left(1+\frac{1}{\tau}\right) \ .$$

5 Constructing Fingerprinting Codes From CEC Codes

Recall we want to distribute slightly different copies of some binary word $v \in \mathbb{F}^n$ to M different users. A fingerprinted copy c must satisfy the following distortion criterion:

$$\mathrm{d}(c, v) \quad \leq \Delta_{\mathrm{o}} \ . \tag{3}$$

The purpose is to trace illicit copies, computed from some fingerprinted copies where "to trace" means to find one of the fingerprinted copy used to create the forgeries. An illicit copy f computed from a set U of original copies must satisfy only a distortion criterion :

$$\min_{c\in U} \quad \mathrm{d}(f, c) \leq \Delta_{\mathrm{f}} \ . \tag{4}$$

If we have a CEC code $\mathcal{C}$ with parameters $(n, M, \Delta_{\mathrm{o}}, 2\Delta_{\mathrm{f}}+1)$, it is possible to solve the fingerprinting problem by using the following scheme: denote by E the encoding mapping of $\mathcal{C}$, and give to the m-th user the fingerprinted copy $E(v, m)$. Criterion (3) is clearly satisfied. Now, if a coalition U creates an illicit copy f then, by (4), there exists m in U such that

$$\mathrm{d}(E(v, m), f) \leq \Delta_{\mathrm{f}} \ .$$

But, $\mathcal{C}$ allows to correct Δ_{f} errors. Thus, decoding f gives $E(v, m)$ and we can recover at least one member of U. Remark that the size of the coalition does not matter, which is a very interesting property.

To be practical, this scheme requires that the code $\mathcal{C}$ has two important properties. On one hand, $\mathcal{C}$ must have an efficient encoding mapping. On the other hand, it must have an efficient decoding algorithm, which is far more restrictive.

From our constructions of centered codes in Corollaries 4.5 and 4.6, we deduce

Theorem 5.1

1. *Let $r, \Delta_{\mathrm{o}}, \Delta_{\mathrm{f}}$ be integers such that $2\Delta_{\mathrm{f}} < \Delta_{\mathrm{o}} < 2^r$. There exist binary $((2^r-1)\Delta_{\mathrm{o}}, 2^{r(\Delta_{\mathrm{o}}-2\Delta_{\mathrm{f}})}, \Delta_{\mathrm{o}})$ fingerprinting codes, with polynomial time encoding and decoding algorithms, which are (∞, e)-secure.*

2. *Let r be an even integer, $\delta_{\mathrm{o}} = (2^r-1)^{-1}$ and δ_{f} be such that $2\delta_{\mathrm{f}} < \delta_{\mathrm{o}}$. Then, for n large enough, there exist $(n, M, n\cdot\delta_{\mathrm{o}})$ fingerprinting codes that are $(\infty, n\cdot\delta_{\mathrm{f}})$-secure, with*

$$\frac{\log M}{n} \gtrsim \left(\delta_{\mathrm{o}} - 2\delta_{\mathrm{f}} - \frac{1}{\left(2^{\frac{r}{2}}-1\right)\left(2^r-1\right)}\right) \log\left(1+\frac{1}{\delta_{\mathrm{o}}}\right) \ .$$

3. *Let $\delta_{\mathrm{o}}, \delta_{\mathrm{f}}$ be real numbers such that $2\delta_{\mathrm{f}} < \delta_{\mathrm{o}} < 1/2$. There exist $(n, M, n\cdot\delta_{\mathrm{o}})$ fingerprinting codes that are $(\infty, n\cdot\delta_{\mathrm{f}})$-secure, with*

$$\frac{\log M}{n} \gtrsim h(\delta_{\mathrm{o}}) - h(2\delta_{\mathrm{f}}) \ .$$

6 On the Capacity Game of Private Fingerprinting Systems

Roughly speaking, the capacity is the highest possible rate of users $\log(M)/n$. For a proper definition, we need to weaken Definition 2.2 of (L, Δ_o)-secure codes to allow some probability of error p_e in the decoding. In this case, we say that the fingerprinting code is $(L, \Delta_f) - p_e$-secure. Now, let n grow to infinity with $\Delta_o = n \cdot \delta_o$ and $\Delta_f = n \cdot \delta_f$. The capacity $C(L, \delta_o, \delta_f)$ is the maximum rate of users $\log(M))/n$ over sequences of $(n, M, n \cdot \delta_o)$ fingerprinting codes, of increasing length, which are $(L, n \cdot \delta_f) - p_e$-secure with p_e decreasing to 0 at infinity.

The model we consider in this paper come from [SBM05], where the capacity of some fingerprinting systems is derived. The fingerprinting systems presented in [SBM05] fulfill two technical assumptions ([SBM05, Def. III.2 and III.3]) : The first one (constant composition) is a constraint on fingerprinting codes, and the second ("smoothness") is a constraint on sequences of fingerprinting codes. A practical consideration is given in order to justify the constant composition assumption: essentially, it allows efficient computation of the fingerprinted copies.

To emphasize the different settings, we denote by $\overline{C}$ the capacity in the sens of [SBM05], that is the capacity restricted to smooth sequences of constant composition codes. Basically, the results are the following:

1. for a fixed coalition size L and the distortions $\Delta_o = n \cdot \delta_o$ and $\Delta_f = n \cdot \delta_f$ growing linearly with the data length n, we have
$$\overline{C}(L, \delta_o, \delta_f) = \mathcal{O}\left(\frac{1}{L}\right) \ ;$$
2. when $L = n \cdot \ell$ grows linearly with n, then
$$\overline{C}(n \cdot \ell, \delta_o, \delta_f) = 0 \ .$$

At first glance, these results seem in contradiction with ours since in Section 5 we construct $(\infty, \delta_f) - 0$-secure fingerprinting codes with an asymptotic rate bounded away from zero. Precisely, Theorem 5.1 states
$$C(\infty, \delta_o, \delta_f) \geq \begin{cases} \left(\delta_o - 2\delta_f - \frac{1}{\left(2^{\frac{r}{2}} - 1\right)\left(2^r - 1\right)}\right) \log\left(1 + \frac{1}{\delta_o}\right) \\ h(\delta_o) - h(2\delta_f) \end{cases}$$
at least for some particular values of the parameters δ_o, δ_f and r. Of course, a possible reason to explain this point is that our construction does not fulfill the two technical assumptions discussed earlier.

At least, we can say that these assumptions have very important consequences and are not so mild as suggested in [SBM05], since they drastically reduce the set of achievable rates. Moreover, they do not lead to fingerprinting codes with more practical properties than our codes, since our codes fulfill the practical considerations, and much more, used to justify the constant composition assumption. Stated in another way, constant composition fingerprinting codes seem to be inefficient.

7 Conclusion

The model we consider in this paper was recently introduced in [SBM05]. This model allows dishonest users to change any part in their own original copies as soon as they don't change too many bits, compared with at least one of their copies.

Recall another well known model, known as the marking assumption [BS98], which allows to change positions in which at least two members of the coalition have different bits. Contrary to intuition, the new model is less favorable to dishonest users than the marking assumption for

which it is not possible to construct binary fingerprinting codes secure against coalitions, even of size two, without allowing some probability of error in the tracing algorithm (see [BS98, Th. IV.2]). Whereas with the model used in [SBM05], we can do it : Namely, we have proved that binary fingerprinting codes without tracing error exist.

In fact, our proof leads to codes with a rate bounded away from zero with a new and interesting property: these codes resist to coalition of arbitrary size. Some heuristic explaining this strange, at first glance, effect is that the condition

$$\exists c \in U, \quad \mathrm{d}(c, f) \leq \Delta_{\mathrm{f}} \ ,$$

is very restrictive, since it means that only a single member of the coalition, namely c, produces a forgery f. Hence, despite that formally the new model deals with coalitions of dishonest users, in fact it reduces to the case of a single user.

By exhibiting very good fingerprinting codes with interesting features, our constructive approach also allows us to prove that the capacity derived in [SBM05, Th. IV.1] for some fingerprinting systems doesn't hold for general ones and that the constant composition assumption is a severe restriction which eliminate, at least asymptotically, all interesting codes.

References

[Bar98] A. Barg. *Handbook of Coding Theory*, volume I, chapter Complexity Issues in Coding Theory, pages 649–754. North-Holland, 1998.

[BBK03] A. Barg, G. R. Blakley, and G. Kabatiansky. Digital fingerprinting codes: Problem statements, constructions, identification of traitors. *IEEE Transactions on Information Theory*, 49(4):852–865, 2003.

[Ber68] E. R. Berlekamp. *Algebraic Coding Theory*. McGraw-Hill, 1968.

[Bli90] V.M. Blinovsky. Asymptotically exact uniform bounds for spectra of cosets of linear codes. *Problems of Information Transmission*, 26(1):83–86, 1990.

[BMP85] G. R. Blakley, C. Meadows, and G. B. Purdy. Fingerprinting long forgiving messages. In *Advances in Cryptology*, number 218 in LNCS, pages 180–189. Springer-Verlag, 1985.

[BP99] L.A. Bassalygo and M.S. Pinsker. Centered error-correcting codes. *Problems of Information Transmission*, 35:30–37, 1999.

[BS98] D. Boneh and J. Show. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[CFNP00] B. Chor, A. Fiat, M. Naor, and B. Pinkas. Tracing traitors. *IEEE Transactions on Information Theory*, 46(3):893–910, May 2000.

[CHLL97] G. Cohen, I. Honkala, S. Listyn, and A. Lobstein. *Covering Codes*. North-Holland, 1997.

[DP86] P. Delsarte and P. Piret. Do most binary linear codes achieve the goblick bound on the covering radius? *IEEE Transactions on Information Theory*, 32(6):826–828, 1986.

[KT74] A.V. Kuznetsov and B.S. Tsybakov. Coding in a memory with defective cells. *Problems of Information Transmission*, 10(2):132–138, 1974.

[McL84] A. McLoughin. The complexity of computing the covering radius of a code. *IEEE Transactions on Information Theory*, 30(6):800–804, 1984.

[MS96] F.J. MacWilliams and N.J.A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 3 edition, 1996.

[SAK+01] K. W. Shum, I. Aleshnikov, P. V. Kumar, H. Stichtenoth, and V. Deolalikar. A low-complexity algorithm for the construction of algebraic-geometric codes better than the Gilbert-Varshamov bound. *IEEE Transactions on Information Theory*, 47(6):2225–2241, 2001.

[SBM05] A. Somekh-Baruch and N. Merhav. On the capacity game of private fingerprinting systems under collusion attacks. *IEEE Transactions on Information Theory*, 51(3):884–899, 2005.

[TV91] M.A. Tsfasman and S. Vlăduţ. *Algebraic Geometric Codes*. Mathematics and its Applications. Kluwer Academic Publishers, 1991.

[TVZ82] M.A. Tsfasman, S.G. Vlăduţ, and T. Zink. Modular curves, Shimura curves and Goppa codes, better than Varshamov-Gilbert bound. *Math. Nachrichten*, 109:21–28, 1982.

[Var97] A. Vardy. The intractability of computing the minimum distance of a code. *IEEE Transactions on Information Theory*, 43(6):1757–1766, 1997.

[vHB76] J.A. van der Host and T. Berger. Complete decoding of triple-error-correcting binary bch codes. *IEEE Transactions on Information Theory*, 22(2), 1976.

[ZC91] G. Zémor and G. Cohen. Error-correcting wom-codes. *IEEE Transactions on Information Theory*, 37(3):730–734, 1991.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399